Scientific Workforce Analysis & Modeling (SWAM)
Project Overview

Submitted to the **National Institutes of Health**, January 2011

Principal Investigators:

Kathryn D. Sullivan, The Ohio State University
Richard C. Larson, Massachusetts Institute of Technology
Michael D. Larsen, The George Washington University
Mary A. Foulkes, The George Washington University

Collaborators:

Joseph Fiksel, The Ohio State University
Josh Hawley, The Ohio State University
Robert H. Tai, University of Virginia

Research Associates:

Emrah Cimren, The Ohio State University
Mauricio Gomez-Diaz, Massachusetts Institute of Technology
Beilei Zhou, The George Washington University
Siyu Qing, The George Washington University
John T. Almarode , University of Virginia
Katherine Dabney , University of Virginia
Devasmita Chakraverty, University of Virginia

Research Objective

This collaborative research project seeks to develop and apply a Scientific Workforce Analysis & Modeling (SWAM) framework for representation and investigation of the diverse and complex forces that influence the quality, availability, and stability of the U.S. academic medicine/health sciences (M/HS) workforce. Building upon previous modeling efforts directed at the field of science technology, engineering and mathematics (STEM), the SWAM framework will consist of a layered set of interrelated dynamic models that will help to evaluate the expected effectiveness and potential unintended consequences of proposed STEM education policies and intervention strategies. Methods for assessing and expressing statistical uncertainty of estimates as well as sensitivity of predictions to modeling assumptions will be included in the model development effort. In addition, to support the modeling effort, the project team will compile and share a growing knowledge base regarding M/HS education and workforce development, providing guidance for future workforce-related research and data collection.

Rationale

Declining graduation rates in STEM disciplines have raised concerns about the availability of a technically-capable workforce to keep the U.S. competitive and secure (National Science Board, 2008). Yet the results of well-meaning interventions have been largely disappointing, arguably because they failed to address underlying systemic

issues. Indeed, the STEM workforce cannot be understood in isolation, since the career pipeline is influenced by the broad context of U.S. innovation, competitiveness, immigration, and other economic and societal factors. NIH has identified several disturbing trends, including rising ages of principal investigators, lower funding rates for female investigators, and persistent educational and career attainment gaps among racial and ethnic minorities. This project will investigate how broader contextual factors interact with specific policies aimed at improving STEM instructional quality, student performance, and retention. Based largely on existing knowledge, it will provide new modeling tools to deepen our understanding of the nation's complex education-workforce "ecosystem" and allow rigorous examination of policy questions, such as:

- How student proficiency interacts with career interest in making educational choices.
- Whether under-represented groups are benefited by various STEM initiatives.
- What factors influence persistence and recruitment to M/HS majors and workforce.

The SWAM framework will permit exploration of a variety of policy scenarios, and will be shared with other parallel research efforts funded by NIH. The long-term goal of this research is to encourage the use of modeling tools by the M/HS community to anticipate future trends, enhance decision-making processes, and identify key research needs.

Research Plan

Under the terms of the Scientific Workforce Cooperative Agreement, Ohio State and its partner universities will model the scientific workforce, focusing specifically on the analysis and modeling of the ***transition from college to the workforce*** for life sciences and/or biological sciences.

The expected result will be a portfolio of models that enable exploratory analysis of the potential impacts of broad categories of policy initiatives, such as educational programs or employment activities, on the long-term research success for individuals engaged in life/biological science careers. The project will be conducted in three phases over a four-year time frame. The following is an overview of the planned activities, and a more detailed description of the plans for Phase 1 is provided below.

Phase 1 (2011). The first phase will consist of baseline model development, including both system dynamics and agent-based modeling approaches. The modeling effort will be informed by interviews with subject matter experts, and harvesting of data sources on M/HS education and workforce patterns. The result will be the establishment of a high-level SWAM architecture and key research hypotheses.

Phase 2 (2012-2013). The second phase will involve extension and refinement of the baseline models. Research will include simulation of specific questions of interest, based on insights from model development and consultation with the NIH Program Officer and other awardees. This phase will also include initial development of user interface tools. Attention will be given to representing uncertainty due to knowledge and data limitations and variation.

Phase 3 (2013-2014). The third and final phase will involve continued work on in-depth policy analysis, model refinement, and user interface implementation. Methods for assessing sensitivity to model and parameter assumptions will be incorporated into products. The team will also develop recommendations for future research and data collection.

During Phase 1, we plan to pursue the following major research thrusts.

- Analyze extant datasets to identify insights that are additive to the literature and that provide a retrospective understanding of student populations and patterns.
- Develop an understanding of the "physics" of decision behaviors regarding college and career choices via guided conversations with a modest target set of people within the system.
- Utilize the results of the above investigations to develop an initial suite of linked or nested models, potentially including statistical, agent-based, system dynamics and Markov chain methods.

Extant Data Sets

The data sets listed in Table 1 can be used to develop an understanding of the scientific workforce educational trajectories. These data sets include:

Table 1. Potential Data Sets to Support Scientific Workforce Analysis & Modeling

Data Set	Agency Source	Sample	Variables of Interest	Access
SESTAT	National Science Foundation	Two sub-samples are of interest (individuals with STEM Occupation in 2006 data; individuals with degrees in STEM fields). Can limit to specific fields of study or work (about 5% biological science). About 100,000 people overall. Three component surveys (Survey of Doctorate Recipients, National Survey of College Graduates, and National Survey of Recent College Graduates) conducted 1993 – 2008.	1) Key data on educational attainment (e.g., field of study, year of degree) 2) Data on employment choices (e.g., working/not; level of effort in work; work-life balance) 3) Data on advancement (e.g., tenure status, rank, management status)	Need restricted use agreement
NLSY (both 1979 and 1997 samples)	Department of Labor/OSU	Two sub-samples of interest (those who major in a STEM field in college; those working in STEM areas). About 12,000 in 1979 sample; few in biological field, but more in STEM.	1) Lots of data on background characteristics, education, occupational choices. 2) Data on employment/un-employment over time	Need restricted use agreement
NELS (1988-1994 and 1988-2000, public-use and restricted use data sets)	National Center for Educational Statistics	Sampling of nationally representative general population tracked longitudinally from Age 14 to Age 26.	Vast array of questions on a variety of students' educational and personal experiences supplemented by responses from parents, teachers, etc.	Restricted-use license held by R.H.Tai, UVA

Institution Specific[1] data on students	Ohio State and potentially others	Sub-samples for analysis could include 1) all graduate students in one or more college (e.g., medicine) or 2) post-doctoral/advanced research students.	Data from this could include both information on educational activities but also the impact of policy specific programs designed to increase engagement in research.	Requires institutional agreements
Project Crossover	NSF-sponsored research study (NSF REC 0440002)	Two data sets (Data Set 1 collected from chemists, physicists, and chemical engineers; Data Set 2 collected from graduate students in chemistry and physics)	Highly focused very specific questions about graduate school experiences and decisions about professional choices and career paths	Proprietary data set held by R.H.Tai, UVA

1. Very large specialized data files like the Scientists and Engineers Statistical Data System (SESTAT) that can model more specific relationships between post-graduate attainment and labor market involvement in research careers,
2. Moderately-sized but high quality national samples (such as the National Longitudinal Survey of Youth and the National Educational Longitudinal Study of 1988) that can be used to generate an understanding of the economy wide high school and college level factors that might impact STEM engagement
3. Small highly focused data sets collected through survey studies examining specific aspects of workforce development such as education and graduate training experiences such as the Project Crossover Survey of Physical Scientists and Graduate Students (NSF REC 0440002, PI R. H. Tai).

It should be noted that institution specific data (e.g., cohort data from Ohio State or another school) could be very useful at obtaining data on micro-interventions that NIH could implement more widely. For instance, Ohio State's College of Medicine has many efforts underway to impact engagement with research on behalf of undergraduate and graduate students. It is possible also to merge with labor market data kept by individual states to obtain long-term economic outcomes for researchers.

Our next steps regarding extant data collection are to:

1. Determine the initial relationships that require statistical analysis (e.g., relationship between high school coursework/GPA in STEM and the decision to major in a STEM field in college; influence of marital status and presence of children by STEM field on the decision to pursue a post doc or assistant professor position).
2. Obtain approval to use SESTAT and other data in its restricted form.
3. Process SESTAT and other data in order to characterize initial relationships of interest. Part of this work could involve associating additional information, such as Carnegie rankings of institutions and summary information from the NSF-NIH Survey of Graduate Students and Post-doctorates in Science and Engineering (also known as the GSS), with data from sources in Table 1.

[1] These are hoped for data sets. We are in the preliminary stages of exploring the possibility of obtaining these data from one or more university.

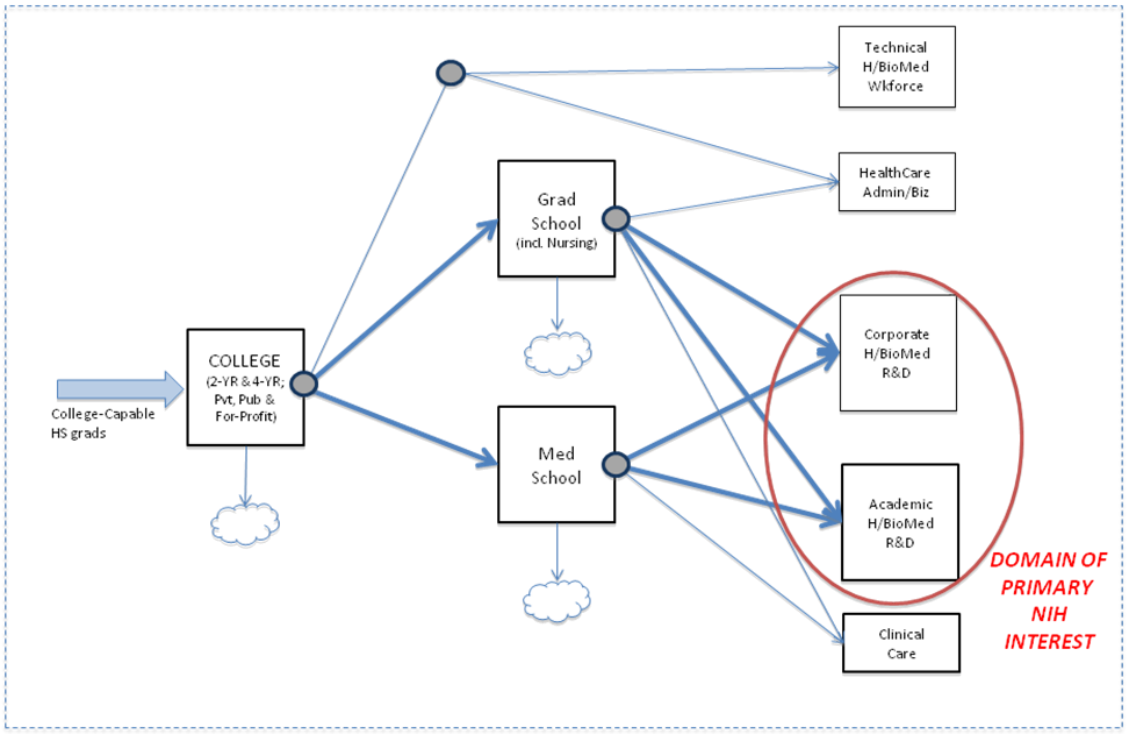
4. Explore the use of institution specific models to understand impact of policies targeted at increasing the long-run engagement of researchers in life/biological sciences.

Of course, one can suggest additional data sources that potentially would be advantageous to examine. A couple of examples illustrate the point when considering careers of doctorate recipients. One might like to examine an individual's grant application and award information from NIH and NSF in relation to factors available on that individual in the above-mentioned data sets. One might also like to have summaries of academic performance, such as might be extracted from CVs or from journal publication citation indexes, when considering issues of promotion, tenure, and persistence in research. Discussion of access to and use of additional data sources certainly would be welcome.

Decision Behaviors

To inform the development of models, we need to understand the "physics" of the process, i.e., the factors influencing decisions by potential young scholars that lead them to continue to pursue STEM careers or divert them to other career paths (as depicted in Figure 1).

Figure 1. College and Career Pathways for Medical/Health Science Students



While we will study these factors in general, we have particular interest in issues concerning women and under-represented minorities. Although the population of interest for this project is STEM-focused, university-admitted students, we may need to revert back to earlier parts of students lives if/when it becomes apparent that causal mechanisms at those earlier stages influence career choice trajectories in college and beyond. Not all paths, of course, are represented in Figure 1. Some individuals enter

graduate or medical school after participating in the technical workforce. There is multidimensional movement among corporate, government, and academic R&D. Other segments of the economy compete for the students that complete M/HS degrees. . In all areas, some individuals take time out of the "pipeline" for family or other pursuits, sometimes re-entering on the same or a modified trajectory. The variety of influences and complexity of possible flows make this a challenging topic.

Modeling Tools

The team will design and create a portfolio of mathematical and statistical models that will shed light on the concerns of the research. These models will focus on causal factors that influence a young person's decision to continue to study STEM careers, some leading to research careers supported in part by the NIH. While the best model choices will only become apparent once we understand the physics and the data available, we expect that the portfolio will include the following:

1. **Cohort-following statistical models** that depict past behaviors and project future trajectories of cohorts of various collections of young people. The cohorts may be characterized by various demographic descriptors, such as age, gender, race, ethnicity, home location in the USA, family income level, marital status, presence of children at home, institution, field of study, and others. Statistical models can include regression model for quantitative outcomes, logistic regression models for estimating probabilities of key transitions, and time to event or "survival" models. Such models are most useful for evaluating historical results. They can be used with caution to predict future behavior only if current policies and incentives remain unchanged, or if we can confidently predict how the cohort behaviors will respond to such changes. For a given statistical model and data set, uncertainty about model parameters and measures of the goodness-of-fit of the model to the data can be calculated. Robustness to model assumptions can be investigated. The statistical models also can be used for simulation.
2. **Agent-based models**, in which a system is modeled as a collection of autonomous decision-making entities, called agents. Each agent individually assesses its situation and makes decisions on the basis of a set of rules. Repetitive competitive interactions between agents are a feature of agent-based modeling, which relies on the power of computers to explore dynamics out of the reach of pure mathematical methods. Even a simple model can exhibit complex behavior patterns and provide valuable information about the dynamics of the real-world system that it emulates. The set of rules and their parameters can be "tuned" to increase the realism of the resulting simulation. Variations in the rules can be used to assess "what if" scenarios.
3. **System dynamics models**, which utilize an intuitive graphical technique for creating aggregate models of systems having non-linearities, feedback loops, and delays. The behaviors that characterize a system can be represented by non-linear differential equations or any arbitrary mathematical functions. Such models are most useful for providing insights and "connecting the dots" in complex systems, including the identification of potential unintended consequences of new policies that appear attractive in the short term. However, the need for aggregation may omit important behavioral issues that are only apparent at the "micro" level.
4. **Markov models** of system movements, which depict the future probabilistic evolution of a system as conditional only on the present state. The definition of state is often one of the creative elements in defining a Markov model. These models are

characterized by a "no memory" property, or limited dependence on stated factors. Even so they have been found useful in a wide variety of settings. Once a system is characterized as a Markov model, there is a rich and deep theory that describes the future behavior of the system. Changes in the Markov model specification can be used to examine sensitivity of results to modeling assumptions.

Undoubtedly, there will be other models surfacing in our portfolio as well, including regression models and decision analysis models. Each will play a role in moving forward our collective understanding of the research questions of the grant. The models will be of interest in and of themselves, but also in relation to one another. For example, the statistical models might inform specification of Markov models, agent-based rules, and system dynamic graphical connections. Similarly, useful agent-based rules and system dynamic connections could suggest new statistical models and new specifications for Markov models.

Limitations and potential obstacles

In a large research effort such as the one proposed here, there naturally will be limitations and potential obstacles. One potential limitation is the availability of data sets and the content of those data sets. No one data set contains all the variables or all the cases that might be informative about existing and past relationships. Phase 3 is included in the research plan, because it is likely that the combined efforts of the research team will identify critical gaps in the available information. In order to address these limitations, the team will develop recommendations for future research and data collection.

A second potential limitation is the limited ability of models to capture the complex relationships that exist in society. It is with this limitation in mind that the teams do not propose examining a single model, but rather a rich set of diverse models to inform the characterization of sample data and support policy investigation. Efforts will be made to express uncertainty, either through statistical measures or simulation, for single models. Sensitivity to model specifications and parameter values also will be examined.

A third potential limitation is the inherent challenge of performing causal analysis in an observational, rather than in an experimental, framework. Indeed, there is no guarantee that observed differences in education, support, and opportunities will translate into realized gains when used to design policies. We will keep this in mind when writing summaries of research findings. To the degree that different data sources and different modeling strategies give concurrent estimates of the impact of policy interventions, one can be more or less certain about the effect range of a given action.

Finally, there is no escaping the fact that some macro level factors, namely the economy (national and international) and behavior of other countries, will continue to affect the condition of the STEM and medicine/health sciences (M/HS) workforces in the U.S. To some degree broad assumptions about these macro level factors will influence some models. Any estimate, however, will be sensitive to large variation in macro level influences. Sensitivity to model specifications and parameter values will give some idea of the uncertainty that should reasonably be associated with stated results.